Improved consistency in estimates of conditional false discovery rates increases power relative to both existing methods and parametric estimators

James Liley, Chris Wallace

Department of Medicine, University of Cambridge

MRC Biostatistics unit, University of Cambridge

ajl88@.cam.ac.uk, cew54@cam.ac.uk

September 27, 2018

Abstract

A common aim in high-dimensional association studies is the identification of the subset of investigated variables associated with a trait of interest. Using association statistics on the same variables for a second related trait can improve power. An important quantity in such analyses is the conditional false-discovery rate (cFDR), the probability of non-association with the trait of interest given p-value thresholds for both traits. The cFDR can be used for hypothesis testing and as a posterior probability in its own right. In this paper, we propose new estimators for the cFDR based on kernel density estimates and mixture-Gaussian models of effect sizes, the latter also allowing estimation of a 'local' form of cFDR (cfdr). We also propose a general non-parametric improvement to existing estimators based on estimating a posterior probability previously estimated at 1. We find that new estimators have the desirable property of smooth rejection regions, but, unexpectedly, do not improve the power of the method, even when distributional assumptions are true. Furthermore, we find that although the local cfdr represents a theoretically optimal decision boundary, noisiness in its estimation means it is less powerful than corresponding cFDR estimates. We find, however, that the non-parametric adjustment increases power for every estimator. We

demonstrate the best method on transcriptome-wide association study datasets for breast and ovarian cancers. The findings from this analysis are of both theoretical and pragmatic interest, giving insight into the nature of cFDR and the behaviour of false-discovery rates in a two-dimensional setting. Our methods allow improved control over the behaviour of the cFDR estimator and improved power in high-dimensional hypothesis testing.

1 Introduction

In the past decade, the progress of biological investigation has been characterised by increasing size of datasets and increasing diversity and precision of phenotypes under investigation. This suggests 'leverage' approaches, in which we seek to learn more about one trait by using data from another. Specifically, given a set of p-values P from association tests for a set of hypotheses for a trait under investigation, and corresponding values Q of some covariate which has different distributions amongst associations and non-associations with P, we seek to use the values of Q to strengthen our ability to find associations with P.

A range of methods for this problem have been proposed. In a Bayesian setting, Ferkingstad *and others* (2008) determine posterior probabilities of association for each hypothesis using a frequentist p-value for the ith hypothesis and a prior probability of association modulated according to the covariate, with values 'binned' into categories. A frequentist approach by Ignatiadis *and others* (2016) weights each hypothesis under consideration using the covariate value for that hypothesis, with a form of the Benjamini-Hochberg procedure used on resultant weighted observations. A third approach by Zablocki *and others* (2014) parametrises the joint distribution of the covariate and p-values allowing continuous modulation of the probability of association with changes in the covariate. This enables estimation of the probability of association for a hypothesis of interest given p-values for both the trait of interest and the covariate, a quantity which we will revisit in this work.

The method we will primarily consider in this paper is termed the conditional false-discovery rate (cFDR). First proposed in Andreassen *and others* (2013), the cFDR represents the probability of non-association with a phenotype of interest (we will denote this event H_0^P) given p-value thresholds p, q on observed p-values P, Q for both the trait under

investigation and a second trait (Q):

$$cFDR(p, q) = Pr(H_0^P | P \leq p, Q \leq q) \tag{1}$$

When estimated for each hypothesis at the observed values of p, q, the cFDR can be used for hypothesis testing. The cFDR has been used successfully to lever secondary datasets in genome-wide association studies (GWAS) (Andreassen *and others*, 2014; Andreasson *and others*, 2014; Liley and Wallace, 2015). The cFDR is powerful in many cases and effective in its simplicity and interpretability.

None of the above methods is clearly dominant in every case. In general, the problem is complex: the optimal method depends on the distribution of effect sizes for the trait under investigation, the type and distribution of covariate, and the number of variables. Furthermore, the best method depends on the desired outcome and measure of type-1 error rate.

In a recent work (Liley and Wallace, 2018), we established a method to control FDR when using cFDR for association testing. While FDR control is assured, the existing method has the shortcoming that rejection regions are often highly discontinuous and dependent on ordering of the p-values. Additionally, estimation of the 'cfdr' statistic used by Zablocki *and others* (2014):

$$cfdr(p, q) = Pr(H_0^P | P = p, Q = q) \tag{2}$$

is not supported. Because our FDR controlling method is immediately adaptable to different estimators of cFDR, there is scope to diversify the range of cFDR associated methods to overcome these shortcomings and adapt to different problems.

In this work, we examine new estimators for cFDR, and evaluate their performance on simulated data. We also find that although the cfdr can be used to construct optimal rejection regions, its estimation is so noisy that it may often be less effective than the cFDR in practice. Finally, we introduce a non-parametric improvement to the existing cFDR method, and show this leads to a substantial improvement in power over existing methods. We apply the best-performing method to datasets from transcriptome-wide association studies (TWAS) as an example.

2 Overview of existing estimator and FDR control

Suppose we have a set of p-value pairs (p_i, q_i) where $p_1, p_2, ..., p_n$ are p-values from association tests in a study of interest considered to be draws from the random variable P, and $q_1, q_2, ..., q_n$ are corresponding p-values from the second study, considered to be draws from Q.

The current estimator cFDR uses a multiset of p-value pairs X, which may or may not be the set S. We expand definition (1) as

$$cFDR(p,q) = \frac{Pr(P \leq p | H_0^P, Q \leq q)}{Pr(P \leq p | Q \leq q)} Pr(H_0^P | Q \leq q) \tag{3}$$

Under the assumption $P \perp\!\!\!\perp Q | H_0^P$, we have $Pr(P \leq p | H_0^P, Q \leq Q) = Pr(P \leq p | H_0^P) = p$. The two remaining quantities (which we will focus on in this paper) are approximated as

$$Pr(P \leq p | Q \leq q) \approx \frac{|\{i : p_i \leq p, q_i \leq q, (p_i, q_i) \in X\}|}{|\{i : q_i \leq q, (p_i, q_i) \in X\}|} \tag{4}$$

and

$$Pr(H_0^P | Q \leq q) \approx 1 \tag{5}$$

giving the estimate

$$c\widehat{FDR}_X^1(p,q) = p \frac{|\{i | p_i \leq p, (p_i, q_i) \in X\}|}{|\{i | p_i \leq p, q_i \leq q, (p_i, q_i) \in X\}|} \tag{6}$$

noting the dependence on X, and noting by the superscript '1' the original estimator. Estimate (4) is generally consistent, and estimate (5) is conservative, meaning that $c\widehat{FDR}_S^1$ is asymptotically conservative in that under reasonably general circumstances

$$\lim_{|S| \to \infty} c\widehat{FDR}_S^1(p,q) - cFDR(p,q) \leq 0 \tag{7}$$

for $(p,q) \in (0,1)^2$.

The FDR-controlling method we use here begins by dividing hypotheses into 'folds'. We then construct 'L-regions' for each observation $(p_i, q_i) \in S$ which roughly represent the set of points (p,q) for which $c\widehat{FDR}(p,q) \leq c\widehat{FDR}(p_i, q_i) = \alpha$. Strictly, we define

$$L_X^\zeta(\alpha) = \{(p,q) : \exists p' \geq p : c\widehat{FDR}_{X+(p',q)}^\zeta(p', q) \leq \alpha\} \tag{8}$$

where ζ indicates the type of cFDR estimator used (so far we have only introduced one estimator in (6) but others will be proposed in Section 3). The L-region for point p_i, q_i is computed leaving out the points in the same fold as p_i, q_i; that is, using $X = S -$ (fold containing i).

The PDF f_0 of $P, Q|H_0^P$ is then estimated, using the fact that $P|H_0^P \sim U(0,1)$ and $P \perp\!\!\!\perp Q|H_0^P$, assuming $Q|H_0^P \approx Q|P > \frac{1}{2}$ and assuming a mixture-Gaussian model for $\Phi^{-1}(Q/2)|H_0^P$ (this assumption is discussed in a later section of this paper). For each such region L, we then define 'v-values' (an analogue of p-values) as

$$v(L) = \iint_A f_0(x,y)dxdy = Pr\left(P, Q \in A|H_0^P\right) \tag{9}$$

and the Benjamini-Hochberg (Benjamini and Hochberg, 1995) method is applied to control FDR. Because the FDR-controlling method uses only regions L rather than actual values of $c\widehat{FD}R$, the values $c\widehat{FD}R$ can be scaled arbitrarily and the results will not change. Importantly, FDR control will be achieved even if the estimate $c\widehat{FD}R$ is non-conservative. For technical details regarding FDR control, we refer the reader to Liley and Wallace (2018).

When used either as a hypothesis test or as a direct estimate, $c\widehat{FD}R_S^1$ can be a poor approximation of the quantity $Pr(P \leq p|Q \leq q)$ in regions near the $P = 0$ and $Q = 0$ boundaries of the unit square. At these extremes, the approximation in (4) has large discontinuities around observed p_i, q_i. This is demonstrated in figure 1, which shows that $c\widehat{FD}R_S^1(p,q)$ can vary twofold in any neighbourhood of certain $p_i, q_i \in S$. Such extreme variation at a small scale is undesirable as very small fluctuations in observed data should not have marked effects on test statistics. It is important to note that a similar problem affects posterior estimates of FDR in the Bayesian sense (Efron *and others*, 2008), and in the Benjamini-Hochberg method itself, in which the acceptance or rejection of any null is dependent on where it falls in the order of other P-values. We propose methods involving parametrisation or smoothing of the bivariate distribution of (P, Q), averting the problem shown in figure 1. We also introduce an approximation of the quantity $Pr(H_0^P|Q \leq q)$ in equation (5), making the estimator of cFDR closer to consistent (up to the factor of $Pr(H_0^P)$, and assuming the consistency of the estimate of $Pr(H_0^P|Q \leq q)$) rather than conservative. We show this improves the power of the method.

3 New metrics for association

3.1 Alternative estimators for cFDR

We now move to consider alternative estimators for cFDR. The estimate $\widehat{cFDR}^1$ is based on empirical quantities estimated directly by counting points (ie, empirical estimation of the joint CDF of (P, Q)) as per definition (6), noting the dependence on the set S. We consider here estimators based on approximating the joint distribution of P, Q: one usinga kernel density estimator (KDE) and one using a bivariate mixture-normal parametrisation. These estimators enforce continuity of $\widehat{cFDR}$ on the open unit square, and thus are robust to small deviations in p-values, overcoming the effect detailed in figure 1. It is easiest to visualise parametrisations as distributions over the unsigned Z scores $(Z_p, Z_q) = \left(-\Phi^{-1}(P/2), -\Phi^{-1}(Q/2)\right)$ with $\Phi^{-1}(x)$ denoting the standard normal quantile function at x.

To avoid distributional assumptions while maintaining a smooth form for the density of P, Q, a second estimator of $Pr(P < p|Q < q)$ can be derived from a two-dimensional kernel density (KDE). We had no reason to prefer any kernel function over another, so opted to use a normal kernel with constant variance I_2. The PDF corresponding to Z_p, Z_q at x, y was modelled in the usual way as

$$g_p(x, y) = \frac{1}{n} \sum_i \frac{1}{\sigma_p \sigma_q} \phi \left(\sqrt{\left(\frac{x - \{-\Phi^{-1}(p_i/2)\}}{\sigma_p} \right)^2 + \left(\frac{y - \{-\Phi^{-1}(q_i/2)\}}{\sigma_q} \right)^2} \right) \tag{10}$$

where $\phi(.)$ is the standard normal density. Values σ_p and σ_q are determined using a standard method based on the observations $p_i, q_i \in X$ (Sheather and Jones, 1991).

We denote the estimator of cFDR derived from the KDE for an arbitrary point (p, q) with Z scores $(z_p, z_q) = (\Phi^{-1}(p/2), \Phi^{-1}(q/2))$ as

$$\begin{aligned} \widehat{cFDR}^2_X(p, q) &= Pr(P \leq p|H_0^P) \frac{1}{Pr\{P \leq p|Q \leq q, (P, Q \text{ have PDF } g_p)\}} \\ &= p \frac{\int_0^\infty \int_{z_q}^\infty g_p(x, y) dx dy}{\int_{z_p}^\infty \int_{z_q}^\infty g_p(x, y) dx dy} \end{aligned} \tag{11}$$

where X in this case refers to the set of points used in fitting the kernel density.

For a third estimator, we use a parametrisation with seven parameters: (π_0, π_1, π_2, τ_1, τ_2, σ_1, σ_2), which parametrise a four-part bivariate mixture-Gaussian distribution over the

$(+,+)$ quadrant with PDF:

$$\begin{aligned} f_p(x,y) &= 4\pi_0 N_{\Sigma_0}(x,y) \\ &+ 4\pi_1 N_{\Sigma_1}(x,y) \\ &+ 4\pi_2 N_{\Sigma_2}(x,y) \\ &+ 4(1-\pi_0-\pi_1-\pi_2) N_{\Sigma_3}(x,y) \end{aligned} \tag{12}$$

where $N_\Sigma(x,y)$ is the PDF of the bivariate normal distribution centred at the origin with variance Σ, the factor of 4 is due to to only unsigned Z-scores being used, and

$$\Sigma_0 = I_2 \qquad \Sigma_1 = \begin{pmatrix} \tau_1^2 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_1^2 \end{pmatrix} \qquad \Sigma_3 = \begin{pmatrix} \tau_2^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \tag{13}$$

This model allows for a proportion π_1 of study variables to be associated only with the trait of interest P (with $SD(Z_P) = \tau_1$), a proportion π_2 to be associated only with the second trait Q (with $SD(Z_Q) = \sigma_1$), and a proportion $(1-\pi_0-\pi_1-\pi_2)$ to be associated with both $(Var(Z_P, Z_Q) = \Sigma_3)$. We allow different values of σ_1, σ_2 and τ_1, τ_2 to allow for potentially different reasons for shared (both P and Q) and independent (P XOR Q) associations. We make the not-generally-true assumption that $Z_P \perp\!\!\!\perp Z_Q$ in the final category in order to keep the model simple and avoid fitting covariances to absolute Z-scores. Maximum-likelihood estimates of parameters can be obtained using an E-M algorithm (Dempster *and others*, 1977). Given fitted parameters $(\hat{\pi_0}, \hat{\pi_1}, \hat{\pi_2}, \hat{\tau_1}, \hat{\tau_2}, \hat{\sigma_1}, \hat{\sigma_2})$, the denominator of (3) can be written:

$$\begin{aligned} Pr(P \le p | Q \le q) &= \frac{Pr(P \le p, Q \le q)}{Pr(Q \le q)} \\ &= \frac{2\int_p^\infty \int_q^\infty f_p(x,y)dxdy}{\int_0^\infty \int_q^\infty f_p(x,y)dxdy} \end{aligned} \tag{14}$$

We denote the estimate of cFDR obtained using this method as

$$\widehat{cFDR}^3_X(p,q) = p\frac{\int_0^\infty \int_{z_q}^\infty f_p(x,y)dxdy}{\int_{z_p}^\infty \int_{z_q}^\infty f_p(x,y)dxdy} \tag{15}$$

where X in this case is the set of points used in the estimation of parameters $(\hat{\pi_0}, \hat{\pi_1}, \hat{\pi_2}, \hat{\tau_1}, \hat{\tau_2}, \hat{\sigma_1}, \hat{\sigma_2})$.

Figure 2 shows the approximate behaviour of L-curves with each cFDR type. Notably,

they are smoothest with $\widehat{cFDR}^3$, although they are constricted to a particular shape.

Estimator $\widehat{cFDR}^3$ models the global distribution of P, Q, allowing simple approximation of the 'local' false-discovery rate:

$$Pr(H_0^P | P = p, Q = q) \tag{16}$$

which can be more useful than the cFDR ($Pr(H_0^P | P \leq p, Q \leq q)$). These quantities cannot be readily estimated by counting points in the manner of $\widehat{cFDR}^1$.

3.2 Local cfdr

An important advantage of the parametrisation approach in $\widehat{cFDR}^3$ is the ability to estimate the quantity $Pr(H_0^P | P = p, Q = q)$ (which we will denote $\widehat{cfdr}$ following the FDR/fdr convention (Efron *and others*, 2008)). We will denote by H_1^P the complement of H_0^P and $f_1(p, q)$ and $f_0(p, q)$ the density functions for $P, Q | H_1^P$ and $P, Q | H_0^P$ respectively.

We show here that the function $cfdr(p, q)$, if known, can be used to generate optimal rejection regions for H_0^P, in the sense of regions R which maximise $Pr(P, Q \in R | H_1^P) = \int_R f_1(p, q) dpdq$ for a fixed value of $Pr(P, Q \in R | H_0^P) = \int_R f_0(p, q) dpdq$ (where H_1^P is the general alternative hypothesis for P). We briefly show that such regions R are regions inside contours of $F(p, q) = \frac{f_1(p,q)}{f_0(p,q)}$.

Let H be the region of the unit square on one side of a contour $F(p, q) = c$, such that for $p, q \in H \Leftrightarrow F(p, q) \geq c$, and suppose $Pr(P, Q \in H | H_0^P) = \int_H f_0(p, q) dpdq = C$ and that $f_0 > 0$ on the interior of the unit square. Let H' be some other region for which $Pr(P, Q \in H' | H_0^P) = \int_{H'} f_0(p, q) dpdq = C$. Now

$$\int_{H' \setminus H} f_0(p, q) dpdq = \int_{H \setminus H'} f_0(p, q) dpdq = C - \int_{H \cap H'} f_0(p, q) dpdq \tag{17}$$

and since $F(p, q) \leq c$ for $(p, q) \in (H' \setminus H) \subseteq H'$, and $F(p, q) \geq c$ for $(p, q) \in (H \setminus H') \subseteq H$,

we have

$$\begin{aligned}\int_{H'} f_1(p,q)dpdq &= \int_{H'\cap H} f_1(p,q)dpdq + \int_{H'\setminus H} f_1(p,q)dpdq \\ &\leq \int_{H'\cap H} f_1(p,q)dpdq + \int_{H'\setminus H} cf_0(p,q)dpdq \\ &= \int_{H\cap H'} f_1(p,q)dpdq + \int_{H\setminus H'} cf_0(p,q)dpdq \\ &\leq \int_{H} f_1(p,q)dpdq \end{aligned} \tag{18}$$

so amongst all regions H' with $Pr(P,Q \in H|H_0^P) = C$ the region H maximises $Pr(P,Q \in H|H_1^P)$. This means that if H is used as a rejection region, it will have the greatest power amongst all rejection regions corresponding to a type-1 error rate of C.

Since

$$\begin{aligned}\frac{1}{cfdr(p,q)} &= \frac{1}{Pr(H_0^P|P=p,Q=q)} \\ &= \frac{f_0(p,q)Pr(H_0^P) + f_1(p,q)Pr(H_1^P)}{f_0(p,q)Pr(H_0^P)} \\ &= 1 + \frac{Pr(H_1^P)}{Pr(H_0^P)}F(p,q) \end{aligned} \tag{19}$$

contours of $F(p,q)$ are also contours of $cfdr(p,q)$. Hence for any test statistic $k(p,q)$ and corresponding threshold α_k for which the rejection procedure $k(p,q) \leq \alpha_k$ controls the type 1 error at α, the test statistic $cfdr(p,q)$ will be the most powerful.

Generally, as $p \to 0$, $q \to 0$, $\Pr(H_0^P|P \leq p, Q \leq q) - Pr(H_0^P|P=p,Q=q) \to 0$, so the value $cFDR(p,q)$ is a reasonably good approximation of $cfdr(p,q)$, and contours of the two functions are similar. However, the argument above suggests that estimators for $cfdr(p,q)$ may outperform estimators for $cFDR(p,q)$ in hypothesis testing. In model 12, null variables (for which $Z_P \sim N(0,1)$) are in the first and third parts, meaning that under the model (writing f_0 as a function of z_p, z_q for convenience)

$$f_0(z_p,z_q) = 4\frac{\pi_0}{\pi_0+\pi_2}N_{\Sigma_0}(z_p,z_q) + 4\frac{\pi_2}{\pi_0+\pi_2}N_{\Sigma_2}(z_p,z_q) \tag{20}$$

allowing the estimator

$$\widehat{cfdr}(p,q) = (\pi_1+\pi_2)\frac{f_0(z_p,z_q)}{f_p(z_p,z_q)} \tag{21}$$

3.3 Estimation of $Pr(Q \leq q|H_0^P)$

The FDR controlling method detailed in Section 2 does not require that $\widehat{cFDR}(p,q)$ be a conservative estimator of $cFDR(p,q) = Pr(H_0^P|P \leq p, Q \leq q)$. We can thus also estimate the quantity (5) as

$$Pr(H_0^P|Q \leq q) = \frac{Pr(Q \leq q|H_0^P)Pr(H_0^P)}{Pr(Q \leq q)} \tag{22}$$

rather than approximating it upward to 1, as done previously. This is useful as the approximation upward to 1 is least accurate (and most conservative) when $Pr(H_0^P|Q \leq q) \ll 1$, and it is in this situation that we can best use the event $Q \leq q$ to help reject $Pr(H_0^P)$. There is no need to estimate the quantity $Pr(H_0^P)$ in the hypothesis-testing setting, as it is constant in all $\widehat{cFDR}$ values and serves only as a scaling factor.

Any of the three estimators $\widehat{cFDR}^1$, $\widehat{cFDR}^3$, $\widehat{cFDR}^2$ can be augmented by including an estimate of the quantity $Pr(H_0^P|Q < q)$, to improve accuracy. We use the estimate

$$\begin{aligned} Pr(H_0^P|Q \leq q) &= \frac{Pr(Q \leq q|H_0^P)}{Pr(Q \leq q)} \\ &\approx \frac{Pr(Q \leq q|P \geq \frac{1}{2})}{Pr(Q \leq q)} \\ &\approx \frac{\left|\{i : q_i \leq q, p_i \geq \frac{1}{2}\}\right|}{\left|\{i : p_i \geq \frac{1}{2}\}\right|} \frac{n}{|\{i : q_i \leq q\}|} \end{aligned} \tag{23}$$

This can be used as a multiplicative factor in any of the estimators above. We denote

$$\widehat{cFDR}_X^{\zeta s} = \widehat{cFDR}_X^{\zeta} \frac{\left|\{i : q_i \leq q, p_i \geq \frac{1}{2}, (p_i, q_i \in X)\}\right|}{|\{i : q_i \leq q, (p_i, q_i \in X)\}|} \frac{|X|}{\left|\{i : p_i \geq \frac{1}{2}, (p_i, q_i \in X)\}\right|} \tag{24}$$

for $\zeta \in \{1, 2, 3\}$ as the 'adjusted' cFDR estimate after inclusion of this factor. Estimation of the density $Pr(Q \leq q|H_0^P)$ is also required for FDR control; namely in estimating the distribution of $P, Q|H_0^P$ in order to integrate it over L. As described in our earlier paper Liley and Wallace (2018) we use a different strategy in this instance, by assuming a parametrisation of the distribution of $Q|P \geq \frac{1}{2}$ and estimating parameters using an expectation-maximisation algorithm. The parametrisation is far simpler to integrate over, so we prefer this method for integrating over regions L. If we used this parametrised estimate of $Pr\left(Q \leq q|P \geq \frac{1}{2}\right)$ in place of the empirical CDF in formula (23), then since the CDF-based estimator of $Pr(Q \leq q)$ is at least $1/n$, the estimate of $Pr(H_0^P|Q \leq q)$ and hence the estimate of $\widehat{cFDR}_X^{\zeta s}(p_i, q_i)$ can be arbitrarily small for low q_i, whatever the value

of p_i, especially if the true distribution of $Q|H_0^P$ is not mixture-normal. We thus prefer the counting-points method for computing the quantity in the context of estimating cFDR values, and the parametrisation for integrating over regions L.

In both the multiplicative factor for $c\widehat{FDR}$ and integration over L, we approximate

$$Pr(Q \leq q|H_0^P) = Pr\left(Q \leq q|P \geq 1/2\right) + \epsilon \tag{25}$$

We note that (assuming $P \perp\!\!\!\perp Q|H_0^P$)

$$\begin{aligned} Pr(Q \leq q|P \geq \frac{1}{2}) &= Pr\left(Q \leq q|P \geq \frac{1}{2}, H_0^P\right) Pr\left(H_0^P|P \geq \frac{1}{2}\right) \\ &\quad + Pr\left(Q \leq q|P \geq \frac{1}{2}, H_1^P\right) Pr\left(H_1^P|P \geq \frac{1}{2}\right) \\ &= \kappa Pr(Q \leq q|H_0^P) + (1-\kappa) Pr\left(Q \leq q|P \geq \frac{1}{2}, H_1^P\right) \end{aligned} \tag{26}$$

where $\kappa = Pr(H_0^P|P \geq 1/2)$, the approximation is consistent in that $\kappa \to 1 \implies \epsilon \to 0$. A larger threshold on P than $1/2$ can and should be used if sufficient points are available to estimate the density on the RHS of equation (25).

We would expect that $\kappa \to 1$ as sample sizes tend to ∞. For fixed finite sample sizes, the sign of ϵ is the sign of

$$Pr(Q \leq q|H_0^P) - Pr\left(Q \leq q|P \geq \frac{1}{2}, H_1^P\right) \tag{27}$$

This quantity is difficult to make any judgements on; a tendency for common effects to both studies to be stronger than associations unique to the conditional trait would increase ϵ, but a tendency for effect sizes to be correlated at these common effects would decrease ϵ.

4 Assessment of performance through simulations

In order to assess the performance of each method above, we simulated a series of datasets and estimated the FDR of each method in each dataset. In each simulation, we generated a set of values p_i, q_i, $i \in 1..n$ from random variables P, Q. We considered an extensive range of underlying parameters governing the distributions of P_i, Q_i. We used an identical simulation protocol to that used in (Liley and Wallace, 2018), reproduced in appendix 1.1. In general, each dataset contained n_1^p associations uniquely in the study P, n_1^q association uniquely in Q,

and n_1^{pq} with both. We used various alternative distributions (normal, t (3df) and Cauchy) for simulating distributions of p-values under the alternative. We selected parameters from continuous distributions where possible to evaluate how power changed continuously.

For each simulation, we defined the 'true detection rate' (TDR) as the proportion of potential true discoveries made:

$$TDR = \begin{cases} 0 & \text{if no rejections} \\ \dfrac{\text{number of correctly rejected null hypotheses}}{\text{total number of associations}} & \text{if} \geq 1 \text{ rejection} \end{cases} \tag{28}$$

so the power to reject H_0^P for a randomly-chosen true association is $E(TDR)$. We controlled the FDR at $\alpha = 0.1$.

We examined how power/TDR varied with the value of $n_1^p + n_1^{pq}$, the total number of variables associated in the study of interest P. This displays the behaviour of estimators over the range of proportions of association variables from no associations at all ($n_1^p + n_1^{pq} = 0$) to a large number of associations.

In these simulations, we assume $Pr(H_0^P|Q \leq q)$ was known for all q for the purposes of integrating over L (that is, when we used the true distribution under which the data were simulated). An analysis of the effect of approximation of $Pr(H_0^P|Q \leq q)$ on FDR is made in Liley and Wallace (2018). All estimates of $Pr(H_0^P|Q \leq q)$ in the computation of $\widehat{cFDR}^{1s}$, $\widehat{cFDR}^{2s}$, $\widehat{cFDR}^{3s}$ were made empirically throughout. We consider the performance of $\widehat{cFDR}^{\zeta}$ with $\zeta \in \{1s, 2, 2s, 3, 3s\}$ compared to $\zeta = 1$.

We compared numerical and graphical analyses of data using paired Wilcoxon rank-sum tests when comparing TDR values between different estimators. In comparing TDR between estimators, all Wilcoxon tests had 90% power to detect a difference between $Pr(\text{TDR with estimator A} > \text{TDR with estimator B})$ and $Pr(\text{TDR with estimator B} > \text{TDR with estimator A})$ of 5% at $p < 0.05$. P-values are not shown but 'difference' refers to $p < 1 \times 10^{-3}$ and 'no difference' refers to $p > 0.1$.

We compared the power of $\widehat{cFDR}^1$, $\widehat{cFDR}^2$, and $\widehat{cFDR}^3$ (figure 3) with each other and with the power of the Benjamini-Hochberg procedure applied to values p_i at the same FDR (which we term BH). Estimator $\widehat{cFDR}^1$ had the highest overall power, followed by estimator $\widehat{cFDR}^3$ and method $\widehat{cFDR}^2$. All method were more powerful than BH. When using a Gaussian alternative distribution, power was similar between estimators $\widehat{cFDR}^1$ and $\widehat{cFDR}^3$, which was unsurprising given the bivariate Gaussian model used in $\widehat{cFDR}^3$.

We found similar results when controlling at a more stringent FDR ($\alpha = 0.01$), shown in supplementary figure 9

In all three estimators, the inclusion of the estimate of $Pr(H_0^P|Q \leq q)$ detailed in Section 3.3 improved the power of the procedure substantially (figure 4)

Power was substantially higher using standard $\widehat{cFDR}^3$ as opposed to $\widehat{cfdr}$, to the extent that $\widehat{cfdr}$ was less powerful for hypothesis-testing than the Benjamini-Hochberg method applied to p-values (figure 5). This was due to the chaoticity of contours derived from cfdr. An example of this is shown in figure 6. FDR control was maintained for all $\widehat{cFDR}$ estimators (data available in github repository in 'Data availability').

5 Conclusions and recommendations

Based on our theoretical and simulated findings, we recommend the use of the $\widehat{cFDR}^{1s}$ estimator for optimal power in hypothesis testing.

We recommend other methods in certain cases. If the alternative distribution of Z scores is likely to be normal (eg, $d = 1$ in table 1) or known, then cFDR estimator $\widehat{cFDR}^{3s}$ (or an equivalent with the known distribution type in place of the Gaussian functions in equation (12)) is of comparable power to $\widehat{cFDR}^{1s}$ (see figure 3, panel 'Normal') and leads to smoother rejection regions and a less chaotic estimator. It also allows computation of $\widehat{cfdr}$ although we do not recommend its use for hypothesis testing unless there is great confidence in the approximation of the distribution of P, Q.

If a conservative estimate of the value $cFDR(p, q) = Pr(H_0^P|P \leq p, Q \leq q)$ is needed, $\widehat{cFDR}^1$, $\widehat{cFDR}^2$, or $\widehat{cFDR}^3$ should be used in place of $\widehat{cFDR}^{1s}$, $\widehat{cFDR}^{2s}$, $\widehat{cFDR}^{3s}$. The inclusion of the estimate of $Pr(H_0^P|Q \leq q)$, which is based on counting points, means that the continuity of $\widehat{cFDR}(p, q)$ on $(p, q) \in (0, 1)^2$ is sacrificed when using $\widehat{cFDR}^{2s}$, $\widehat{cFDR}^{3s}$ in place of $\widehat{cFDR}^2$, $\widehat{cFDR}^3$. Hence if continuity of the estimator in p, q is required, $\widehat{cFDR}^2$ or $\widehat{cFDR}^3$ should be used rather than $\widehat{cFDR}^{2s}$ or $\widehat{cFDR}^{3s}$.

6 Demonstration of method on TWAS data

In this section, we demonstrate the performance of the cFDR method on a real-world dataset from two TWAS. TWAS is a method to attempt discovery of gene-trait associations, as opposed to the variant-trait associations found by genome-wide association studies (GWAS).

A TWAS firstly uses GWAS and expression-quantitative trait locus (eQTL) data to predict mRNA expression levels in a given tissue for each individual, then compares these predicted expression levels across a trait of interest Gusev *and others* (2016).

We considered TWAS datasets for breast cancer (BRCA, Michailidou *and others* (2017)) and ovarian cancer (OCA, Phelan *and others* (2017)), containing tests for varying numbers of genes across 54 tissues. BRCA and OCA have considerable phenotypic overlap Greene *and others* (1984), and we may hope that summary statistics for one disease may be useful for leverage in association analyses of the other. We considered RNA-tissue pairs available in both datasets, restricting our analysis only to pairs in which RNA expression was predicted using data from the GTEx consortium, comprising a total of $n = 80222$ hypotheses.

Given the GWAS-scale dimensionality of testing, we chose a conservative FDR control level $\alpha = 1 \times 10^{-6}$. We used $\widehat{cFDR}^{1s}$ for cFDR estimation (and $\widehat{cFDR}^{1}$ for contrast). We assigned folds according to genes, so expression levels for each gene were assigned a separate fold (for 11327 folds in total). Figure 7 shows z-scores for BRCA and OCA, and rejection regions for cFDR and p-value at the same level of FDR control α. The analysis of BRCA conditioning on OCA enabled 38 more gene-tissue association discoveries than when the p-value alone was used and notably nine more discoveries than when $\widehat{cFDR}^{1}$ was used (724 for $\widehat{cFDR}^{1s}$ vs 715 for $\widehat{cFDR}^{1}$ vs 678 for p-value). In the corresponding analysis for OCA conditioned on BRCA, both $\widehat{cFDR}^{1s}$ and $\widehat{cFDR}^{1}$ enabled four more discoveries than when using p-values alone (310 vs 306). Subjectively, rejection regions from $\widehat{cFDR}^{1s}$ appear more natural than those from $\widehat{cFDR}^{1}$ or from p-values alone, adapting to the joint distribution of Z-scores without requiring any prior hypothesis of joint association.

We found that the relative number of discoveries with the two methods varied considerably with α (figure 8). This observation is specific to this application, and should not be generalised to other TWAS nor to any assumptions on the behaviour of the cFDR. We also do not endorse the universal use of any FDR threshold (including the one used above) in TWAS, and especially note that α should not be chosen to maximise the number of discoveries after an analysis such as that in figure 8. A study-appropriate threshold should be determined *a priori* by consideration of the relative cost of type 1 and 2 errors.

This section is intended to be demonstrative rather than exploratory and we do not claim that putative cFDR-based discoveries are novel. Since this demonstration only involves a single dataset, any judgement about the appropriateness of cFDR to a particular application

should be made using the results of the simulations above, rather than the performance on this dataset.

7 Discussion

There are two main ideas in this work; firstly, the use of parametric or KDE-based estimators to address the discontinuity exhibited by the empirical estimator $\widehat{cFDR}^1$ and secondly the augmentation of any estimator by inclusion of an estimate of $Pr(H_0^P|Q \leq q)$. While they were generally less powerful than $\widehat{cFDR}^1$, parametric and KDE-based cFDR both reduced the chaoticity of the estimator around extremal points. An important pragmatic finding of this work was that the standard non-parametric cFDR is hard to improve on. We showed this both in a theoretical sense, in that it resembles the optimal cfdr, and in an empirical sense in the observation that neither of the alternative estimators were stronger.

The parametric cFDR estimator $\widehat{cFDR}^3$ was slightly more powerful than the non-parametric estimator $\widehat{cFDR}^1$ when the underlying distributional assumption (Normal alternative) was correct. The difference in power in this case was small, and $\widehat{cFDR}^1$ was considerably more powerful when the distributional assumption was false. As $n \to \infty$, we expect that $\widehat{cFDR}^1$ will be more powerful than $\widehat{cFDR}^3$ if distributional assumptions are false, and both will have equivalent power if they are true (since L-regions associated with $\widehat{cFDR}^1$ will converge towards the 'true' L-regions associated with the underlying P, Q distribution, and L-regions associated with $\widehat{cFDR}^3$ will converge towards those associated with the best normal approximation to it).

The parametric cFDR estimator has several theoretical advantages. First, it invites an extension to higher-dimensional spaces, conditioning on two traits rather than one, which is an interesting avenue for future work. Second, it allows computation of the 'local' cfdr, which defines a different ordering on hypotheses than the cFDR. This contrasts with the one-dimensional case, where fdr estimates, FDR estimates, and p-values are generally monotonic (Efron *and others*, 2008). Although the cfdr corresponds to a theoretically optimal decision boundary for rejecting H_0^P if estimated accurately, we found that the estimate was too chaotic to be of general use in hypothesis testing with $n < 10^4$.

It is possible that other estimators of cfdr may be more stable. However, we suspect that similar problems are likely to be encountered, as PDFs are generally harder to estimate than CDFs, even in the univariate setting (Efron *and others*, 2008); heuristically, to estimate the

PDF of P, Q at a point (p, q), either only the small proportion of the data local to (p, q) can be used, or strong parametric assumptions must be made to allow distant points to affect the estimate. We note that cfdr performed poorly even when distributional assumptions were correct and the underlying distribution was well-behaved (Normal) and had only seven degrees of freedom. These artificial conditions are far simpler than those encountered in 'real-world' problems, and we conclude that while the cfdr may remain a useful Bayesian quantity in assessing the posterior probability of association for an interesting variable, no estimator of cfdr is likely to be of use in hypothesis-testing at these scales ($n < 10^4$). Other estimators (for instance, that of Zablocki *and others* (2014)) may be effective at higher n.

Inclusion of an estimate of the quantity $Pr(H_0^P|Q \leq q)$ in the cFDR estimate improves the power over existing methods, and the accuracy of the estimate, at the cost of asymptotic conservatism. The assumption $H_0^P \Leftrightarrow P \geq \frac{1}{2}$ is not ideal, since the sign of ϵ in equation (25) cannot usually be confidently determined, and the estimation is not consistent as $n \to \infty$ (although it is if sample sizes $\to \infty$). Estimates of $Pr(Q \leq q|H_0^P)$ may thus be biased. Alternative methods for estimating $Q|H_0^P$ may include examining the distribution of Q amongst P-values reaching a higher threshold on P (for instance, assuming $H_0^P \Leftrightarrow P \geq 0.99$) or amongst a set of 'control' variables known to be in H_0^P and typical of all variables in H_0^P. However, in Liley and Wallace (2018), we found that the estimate tended to be adequate in the context of integrating over regions L.

In summary, our methods add to the body of methods for co-analysis of omics data. Estimating $Pr(H_0^P|Q \leq q)$ improves the power of cFDR estimators, and parametric estimators enable greater flexibility in their use. We present empirical arguments that although the cFDR is not a theoretically optimal discriminator, it may be pragmatically the best available. These results support confidence in the robustness of the $\widehat{cFDR}^1$ estimator generally and associated FDR-controlling procedures, even at relatively small values of n.

8 Software

Software is available as an R package at `https://github.com/jamesliley/cfdr`. All data and a pipeline to deterministically generate all plots and results in this paper can be found at `https://github.com/jamesliley/cfdr_estimation_pipeline`

9 Supplementary Material

1. Description of our protocol for simulations

2. Supplementary figure: comparison of cFDR estimators at $\alpha = 0.01$

Acknowledgments

CW is funded by the Wellcome Trust (WT107881) and the MRC (MC_UU_00002/4). JL is funded under by WT107881 and for part of this work was on the Wellcome Trust PhD programme in Mathematical Genomics and Medicine, funded by the NIHR Cambridge BRC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: None declared.

References

Andreassen, Ole A., McEvoy, Linda K, Thompson, Wesley K, Wang, Yunpeng, Reppe, Sjur, Schork, Andrew J, Zuber, Verena, Barrett-Connor, Elizabeth, Gautvik, Kaare and Aukrust, Pal. (2014, April). Identifying common genetic variants in blood pressure due to polygenic pleiotropy with associated phenotypes. *Hypertension* **63**.

Andreassen, Ole A., Thompson, Wesley K, Schork, Andrew J, Ripke, Steven, Mattingsdal, Morten, Kelsoe, John R, Kendler, Kenneth S, O'Donovan, Michael C, Rujescu, Dan, Werge, Thomas *and others*. (2013, April). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLOS Genetics* **9(4)**.

Andreasson, Ole A, Harbo, H F, Wang, Yunpeng, Thompson, Wesley K, Schork, Andrew J, Mattingsdal, Morten, Zuber, Verena, Bettella, F and Ripke, Stephan. (2014). Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. *Molecular psychiatry*, 1–8.

Benjamini, Yoav and Hochberg, Yosef. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, series B (methodological)* **57**(1), 289–300.

Dempster, Arthur P, Laird, Nan M and Rubin, Donald B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B (methodological)* **39**(1), 1–38.

Efron, Bradley *and others*. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**(1), 1–22.

Ferkingstad, Egil, Frigessi, Arnoldo, Rue, Håvard, Thorleifsson, Gudmar, Kong, Augustine *and others*. (2008). Unsupervised empirical bayesian multiple testing with external covariates. *The Annals of Applied Statistics* **2**(2), 714–735.

Greene, Mark H, Clark, Jeffrey W and Blayney, Douglas W. (1984). The epidemiology of ovarian cancer. In: *Seminars in oncology*, Volume 11. Elsevier. pp. 209–226.

Gusev, Alexander, Ko, Arthur, Shi, Huwenbo, Bhatia, Gaurav, Chung, Wonil, Penninx, Brenda WJH, Jansen, Rick, De Geus, Eco JC, Boomsma, Dorret I, Wright, Fred A *and others*. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48**(3), 245.

Ignatiadis, Nikolaos, Klaus, Bernd, Zaugg, Judith B and Huber, Wolfgang. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods* **13**(7), 577.

Liley, James and Wallace, Chris. (2015). A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics. *PLOS Genetics*.

Liley, James and Wallace, Chris. (2018). Accurate error control in high dimensional association studies using conditional false discovery rates. *BiorXiv*.

Michailidou, Kyriaki, Lindström, Sara, Dennis, Joe, Beesley, Jonathan, Hui, Shirley, Kar, Siddhartha, Lemaçon, Audrey, Soucy, Penny, Glubb, Dylan,

Rostamianfar, Asha *and others*. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**(7678), 92.

Phelan, Catherine M, Kuchenbaecker, Karoline B, Tyrer, Jonathan P, Kar, Siddhartha P, Lawrenson, Kate, Winham, Stacey J, Dennis, Joe, Pirie, Ailith, Riggan, Marjorie J, Chornokur, Ganna *and others*. (2017). Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nature genetics* **49**(5), 680.

Sheather, Simon J and Jones, Michael C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 683–690.

Zablocki, Rong W, Schork, Andrew J, Levine, Richard A, Andreassen, Ole A, Dale, Anders M and Thompson, Wesley K. (2014). Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics* **30**(15), 2098–2104.

General notes for notation

1. P, p refer to study under investigation, Q, q to study for conditioning on. P and Q are also used to refer to these two studies as opposed to the random variables corresponding to p-values.

2. i and j index hypotheses and variables corresponding to hypotheses, (i) generally meaning the ith smallest

3. S refers to the set $(p_1, q_1), (p_2, q_2), ..., (p_n, q_n)$

4. k indexes folds/subdivisions

5. n is the total number of hypotheses

6. N is the number of folds for cross-validation/resampling

7. P, Q are random variables, p, q are observations

8. H^P is an indicator variable for association in study P. It can be considered in a Bayesian sense as a Bernoulli-distributed random variable, or a frequentist sense as a null hypothesis eg $H^P = 0$.

9. v-values refer to values $v(L)$ used for testing in some sense. The exact definition depends on the method used for FDR control.

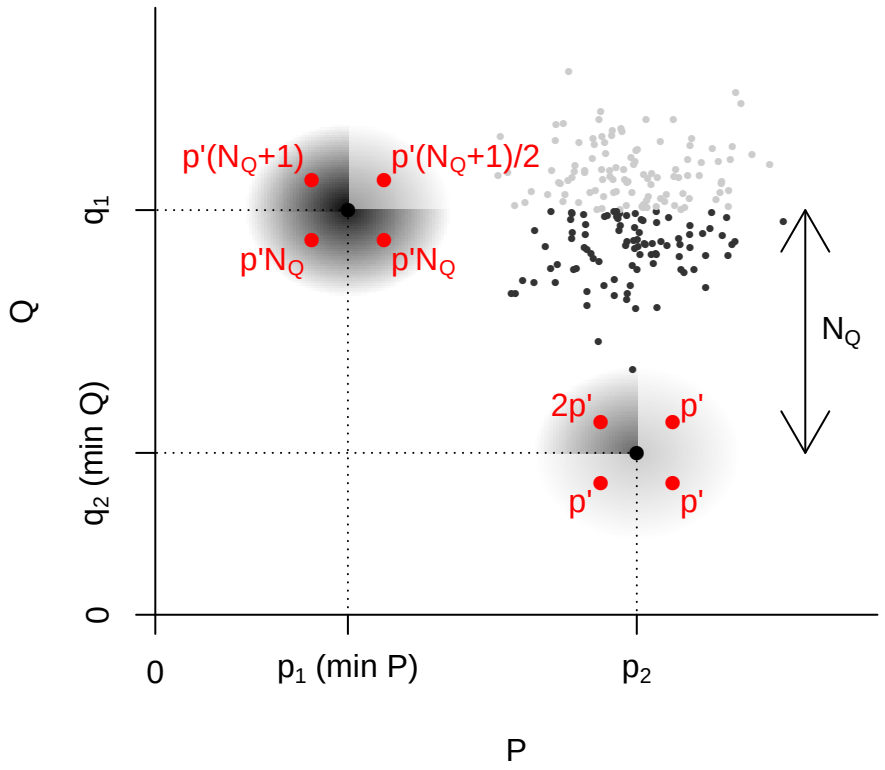

Figure 1: Dependence of $c\widehat{FDR}^1$ values on location of nearby points. In this example, we denote by (p_1, q_1), (p_2, q_2) the points at the (unique) left and lower extremes of the observed p-value distribution respectively; that is, $p_1 = \min(p_i)$, $q_2 = \min(q_i)$. We set N_Q as the number of points with $q_1 \leq q_i \leq q_2$ (small black points). If we add a test point (p', q') (shown in red) in a small neighbourhood of either (p_1, q_1) or (p_2, q_2), the estimated $c\widehat{FDR}^1_{S+(p',q')}(p', q')$ (shown in red next to the point) differs by a factor of 2 in different quadrants of the neighbourhood.

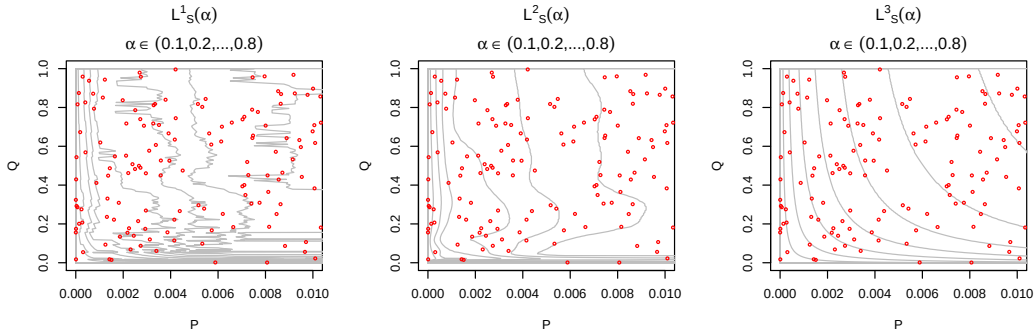

Figure 2: Forms of regions $L_S^{\zeta}(\alpha)$ for $\zeta \in 1, 2, 3$ (CDF, KDE, and modelled estimator respectively) for various values of α. Gray curves show rightmost border of regions, and red circles show datapoints in S. Curves (and hence rejection regions) are smoothest when $\zeta = 2$, but are least responsive to local changes in data density. All regions tend to widen with lower Q.

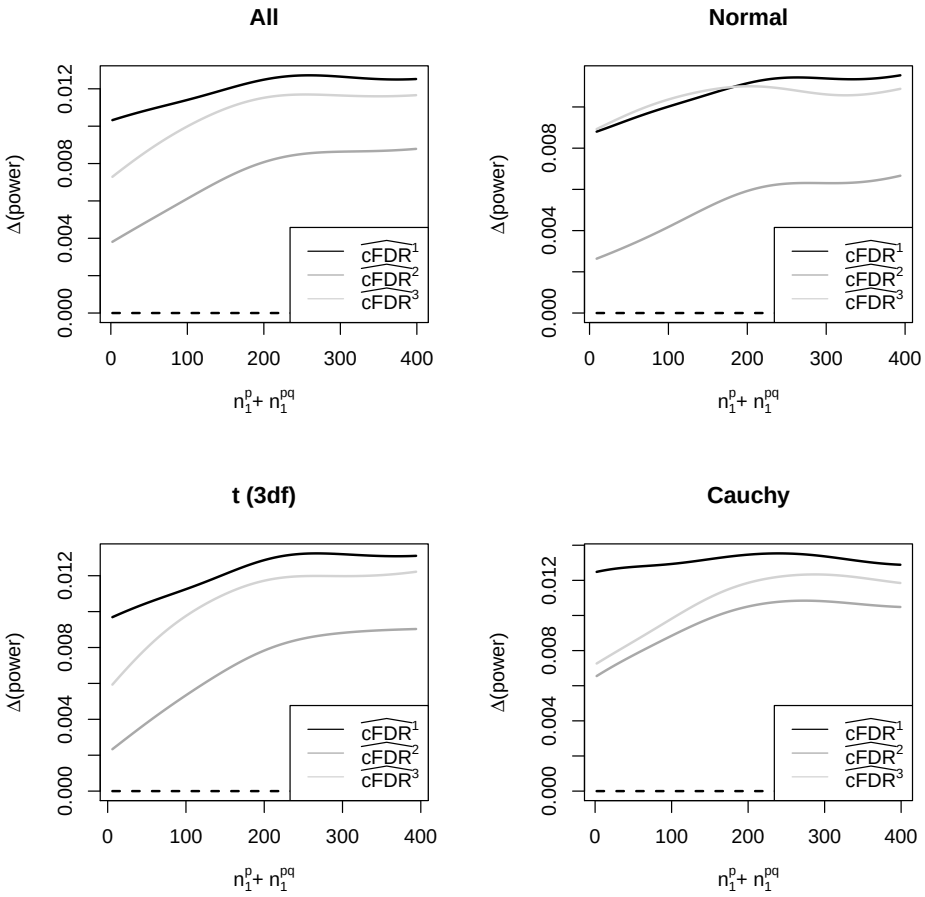

Figure 3: Relative power of each cFDR type ($\widehat{cFDR}^1$, $\widehat{cFDR}^3$, and $\widehat{cFDR}^2$) compared to BH, plotted against $n_1^p + n_1^{pq}$ and subdivided by alternative distribution type (normal, t (3 df), Cauchy, and all combined with frequency 1/3). Parameters drawn from continuous distributions (see appendix 1.1. Gaussian smoothing is used with a kernel width of 1/8 of the x-axis range.

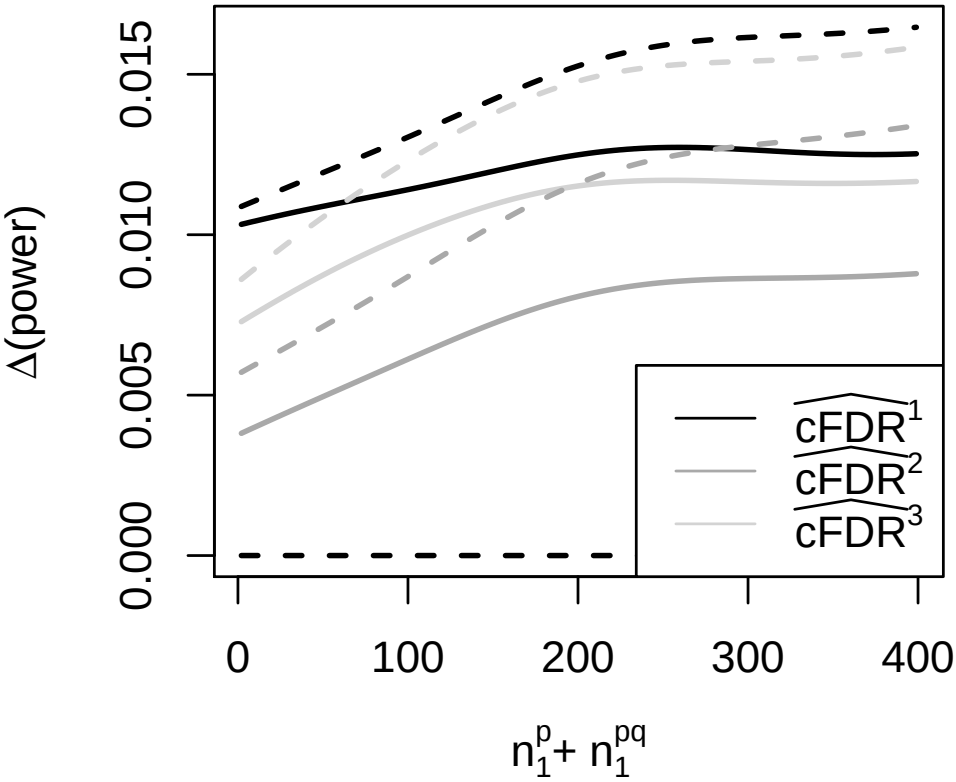

Figure 4: Difference between power of each cFDR type ($\widehat{cFDR}^1$, $\widehat{cFDR}^2$, and $\widehat{cFDR}^3$) with and without inclusion of the estimate of $Pr(H_0^P|Q \leq q)$ (that is, $\widehat{cFDR}^{1s}$, $\widehat{cFDR}^{2s}$, and $\widehat{cFDR}^{3s}$), all compared to BH (the power of the Benjamini-Hochberg method applied to values p_i. Dashed lines show Δ(power) for $\widehat{cFDR}^{1s}$, $\widehat{cFDR}^{2s}$, and $\widehat{cFDR}^{3s}$.

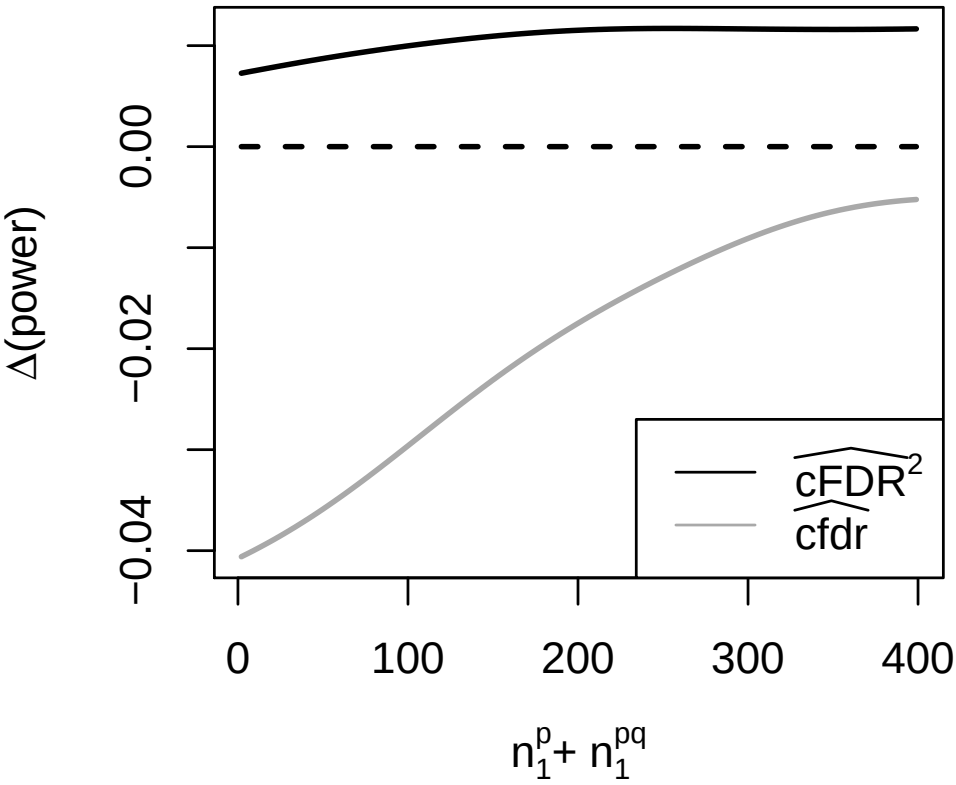

Figure 5: Power of local cfdr estimate $\widehat{cfdr}$ compared to cFDR estimate $\widehat{cFDR}^3$, relative to the power of the Benjamini-Hochberg method applied to p-values. The local cfdr estimate performs poorly.

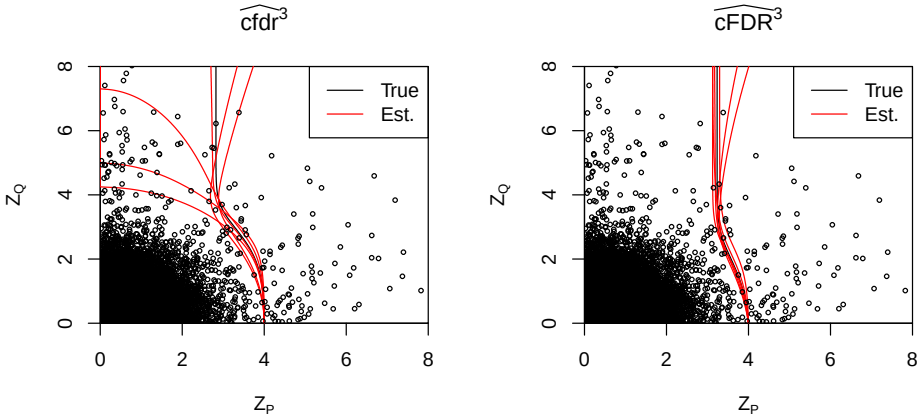

Figure 6: Contours of $\widehat{cfdr}^3$ ('local' PDF-based estimator) and $\widehat{cFDR}^3$ (standard CDF-based estimator) evaluated on the same data. Data were repeatedly simulated (six times) under the true parameter set $(\pi_0, \pi_1, \pi_2, \sigma_1, \sigma_2, \tau_1, \tau_2) = (0.94, 0.02, 0.02, 3, 3, 3, 3)$, with $n = 10^5$ variables. For each simulation, MLE parameters were re-estimated. Red curves show contours of $\widehat{cFDR}^3/\widehat{cFDR}^3$ passing through $(Z_P, Z_Q) = (4, 0)$, and black curves contours derived from the true (rather than estimated) parameters. Contours for $\widehat{cfdr}^3$ are far more chaotic, leading to loss of power despite theoretical optimality. Black points show a realisation of simulated Z_P, Z_Q.

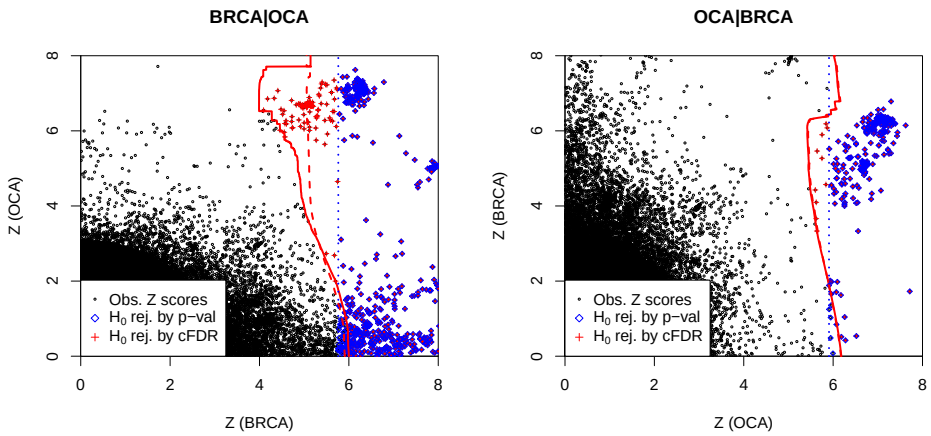

Figure 7: Z scores from TWAS on breast cancer (BRCA) and ovarian cancer (OCA) for a range of genes in 53 tissue types. Red points (crosses) show rejections using $c\widehat{FDR}^{1s}$, blue points (diamonds) using p-value alone, controlling overall FDR at $\alpha = 1 \times 10^{-6}$. Red (solid) and blue (dashed) lines indicate approximate rejection regions for $c\widehat{FDR}^{1s}$ and p-values respectively. Dashed red lines (underneath solid lines on right panel) denote approximate rejection regions for $c\widehat{FDR}^{1}$

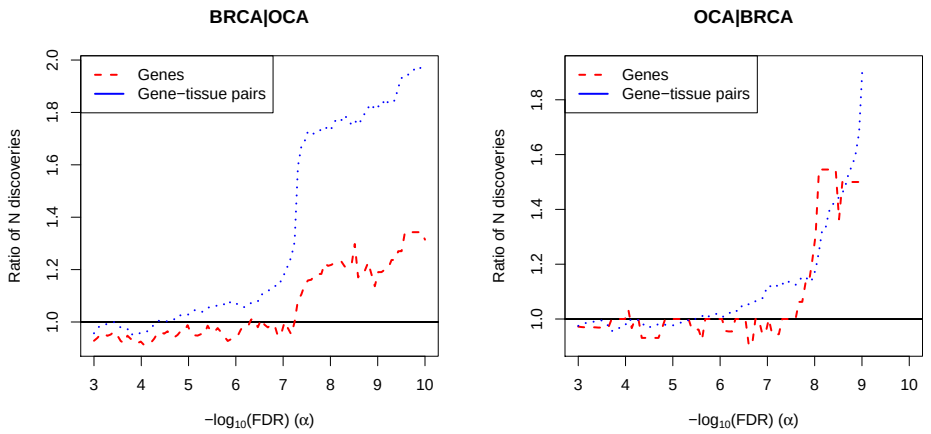

Figure 8: Ratio of new discoveries using $c\widehat{FDR}^{1s}$ and new discoveries using p-value at varying α. New gene-disease associations are shown in red, new gene-tissue-disease associations in blue. Both trend upwards with decreasing α.

1 Appendix

1.1 Simulation protocol

The distributions and variables we considered are shown in table 1. In general, our simulations took the form of: a fixed number n_1^{pq} of variables associated in both P and Q, a fixed number of variables n_1^p associated only with p, a fixed number of variables n_1^q associated only with Q and the remainder of variables associated with neither P nor Q. We will refer to these four classes of variables as C_1, C_2, C_3, C_4. We designated that within each class of variables, P_i and Q_i were iid.

Table 1: Variables used in simulations

Variable	Description	Point values	Sampling distribution for other values
n	Total number of variables	10^3, 10^4	$10^{U(3,4)}$ (rounded)
n_1^{pq}	Number of variables assoc. with P, Q	0, 10, 200	$U(0, 200)$
n_1^p	Number of variables associated with P	0,10,200	$U(0, 200)$ (rounded)
n_1^q	Number of variables associated with Q	0,10,200	$U(0, 200)$ (rounded)
s_p	Scale for distribution of P_i (see below)	$\frac{3}{2}$, 3	$U\left(\frac{3}{2}, 3\right)$
s_q	Scale for distribution of P_j	$\frac{3}{2}$, 3	$U\left(\frac{3}{2}, 3\right)$
d	Form of distributions	Normal, t (3df), Cauchy (eq. prob.)	Normal, t (3df), Cauchy (eq. prob.)

We considered two types of simulations: either with all numerical parameters drawn uniformly from a set of fixed 'interesting' values, or with all numerical parameters drawn from continuous distributions. Our intent was to evaluate FDR values accurately at these 'interesting' points in the parameter space, and more roughly consider values in between these points. One such 'interesting' area of the parameter space was the scenario in which no variables were associated in study P (that is, $n_1^p + n_1^{pq} = 0$). In some plots (namely when looking at how power varies with $n_1^p + n_1^{pq}$) only simulations with parameters drawn from continuous distributions are used.

For variables in C_1, C_2, we set the distribution of P_i (determined by d, s_p) by first simulating Z scores:

d=1: $-\Phi^{-1}\left(\frac{P_i}{2}\right)\frac{1}{s_p} \sim N(0,1)$

d=2: $-\Phi^{-1}\left(\frac{P_i}{2}\right)\frac{1}{s_p} \sim t(\text{df} = 3, \text{ncp} = 0)$

d=3: $-\Phi^{-1}\left(\frac{P_i}{2}\right)\frac{1}{s_p} \sim \text{Cauchy}(\text{location} = 0, \text{scale} = 1)$

where $-\Phi^{-1}\left(\frac{P_i}{2}\right)$ can be considered a Z-score corresponding to P_i, and s_p a scaling factor for the distribution. We set the distribution of Q_i in C_1, C_3 similarly, with s_q in place of s_p. The values p_i, q_i for $i \in C_4$ were sampled from $U(0, 1)$.

Within each class $C_1 - C_4$ we simulated independent P, Q, although P, Q are still dependent when not conditioning on class. We also considered s_p to be constant across C_1 and C_2, s_q to be constant across C_1 and C_3, and d to be constant in each simulation.

We analysed each simulated dataset in parallel using each of the three estimators of $c\widehat{FDR}$. We also performed a standard Benjamini-Hochberg procedure on the values of P (without considering the values of Q) as a control. We further considered each estimator of $c\widehat{FDR}$ both with and without the adjustment due to estimation of $Pr(H_0^P|Q \leq q)$. To estimate the null distribution of $Pr(H_0^P|Q \leq q)$ in order to integrate to generate values $v(L)$, we considered both the true distribution and a mixture-normal distribution estimated from the distribution of $P, Q|P \geq \frac{1}{2}$ as per Section 3.3.

2 Supplementary figure: comparison of cFDR estimators at $\alpha = 0.01$

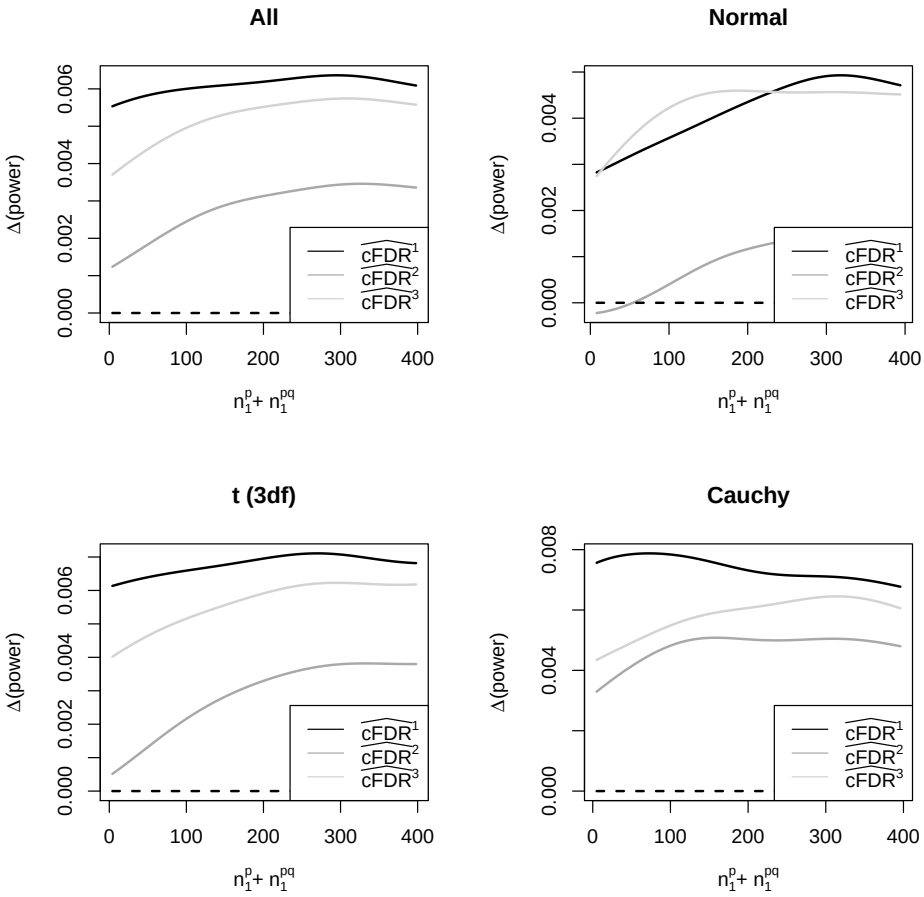

Figure 9: Analogous to figure 3 with $\alpha = 0.01$ instead of $\alpha = 0.1$. Relative power of each cFDR type ($\widehat{cFDR}^1$, $\widehat{cFDR}^3$, and $\widehat{cFDR}^2$) compared to power of the Benjamini-Hochberg method, using FDR control method 3b, plotted against $n_1^p + n_1^{pq}$ and subdivided by alternative distribution type (normal, t (3 df), Cauchy, and all combined with frequency 1/3). Parameters drawn from continuous distributions (see appendix 1.1). Gaussian smoothing is used with a kernel width of 60.